CLAIMS

WHAT IS CLAIMED IS:

1. A system for creating an aggregated data model from a plurality data distribution models, each data distribution model describing a data distribution having one or more data elements, each data element having a value, each data distribution model having one or more bins, each bin comprising a start point having a value, an end point having a value, a value indicating the number of data elements for each bin, and a polynomial formula associated with each bin, the polynomial formula approximating the data elements for the respective bin, comprising:

a processor; and

a computer program executable on a processor, the computer program adapted to perform the following steps:

(a) determining which start point has the minimum value and which end point has the maximum value of all of the bins of all of the data distribution models;

(b) setting a start point of a first bin of the aggregated data model to said start point determined to have the minimum value;

(c) setting an end point of a last bin of the aggregated data model to said end point determined to have the maximum value;

(d) determining a total number of a plurality of points for the aggregated data model by adding the values indicating the number of data elements from all bins from all data distribution models;

(e) approximating the data elements in the data distribution described by each data distribution model using the start point, polynomial formula, and number of data elements for each bin in each respective data distribution

model, each approximated data element comprising one of said points in the aggregated data model;

(f) sorting the points from minimum to maximum;

(g) distributing the points into one or more bins in the aggregated data model such that a substantially equal number of points are in each bin of the aggregated data model; and

(h) determining a polynomial formula with the sorted data elements for each bin of the aggregated data model.

2. The system of claim 1, wherein the computer program is further for determining the end point for each bin in the aggregated data model.

3. The system of claim 1, wherein the computer program is adapted to perform the step of distributing the points into the one or more bins of the aggregated data model according to the following formula::

(a) if the number of points in the aggregated data model is equally divisible into the number of bins, the end point of the first bin is equal to the value of the ith point in the aggregated data model, wherein i is the number of points in each bin determined by dividing the points equally into the number of bins, wherein the value of the end point of each bin is equal ith point after the last point in the proceeding bin, wherein the start point of each bin is equal to the point after the last point of the previous bin, else

(b) if the number of data elements in the points is not equally divisible by the number of bins, then the number of points in each bin is determined by dividing the number of points by the number of bins, and then adding one to the count of the points in each of a number of bins equal to the

remainder after dividing, wherein the bins that have one added to the count is determined according to the following formula:

for k from 1 to r

$$bin_{add} = INT((n*k) / (r + 1))$$

next k

wherein bin_{add} is the sequential bin number to add one to the count of points to include therein, n is the total number of bins in the aggregated data model, r is the remainder from dividing the number of points in the data distribution by the number of bins, and INT is a function for rounding the result of the bracketed formula to produce an integer result.

4. The system of claim 1, wherein the computer program is for performing separately for each bin of the aggregated data model, the steps of approximating the data elements for each bin, determining the end point for each bin, and determining the polynomial formula for each bin.

5. The system of claim 1, wherein each data distribution model is the result of the computer program performing a the following steps:

(a) sorting the data elements in from minimum to maximum for each data distribution;

(b) computing the number of data elements in each data distribution;

(c) determining the value of the start point and the value of the end point of each bin by dividing the data elements into a plurality of substantially equal sized bins for each data distribution;

(d) counting the number of data elements in each bin for each data distribution; and

(e) computing each distribution model for each data distribution, each distribution model comprising, for each bin, the start point of the bin, the end point of the bin, and the number of data elements in the bin.

6. The system of claim 5, wherein the computer program is adapted to perform the following steps for determining the start points and end points of the bins for each data distribution model:

selecting as the start point of the first bin the value of the data element having the minimum value in the sorted data distribution;

determining the start point and end point of each bin according to the following criteria:

(c) if the number of data elements in the data distribution is equally divisible into the number of bins, the end point of the first bin is equal to the value of the ith data element in the data distribution, wherein i is the number of data elements in each bin determined by dividing the data elements equally into the number of bins, wherein the value of the end point of each bin is equal ith data element after the last data element in the proceeding bin, wherein the start point of each bin is equal to the data element after the last data element of the previous bin, else

(d) if the number of data elements in the data distribution is not equally divisible by the number of bins, then the number of data elements in each bin is determined by dividing the number of data elements by the number of bins, and then adding one to the count of the data elements in each of a number of bins equal to the remainder after dividing, wherein the bins that have one added to the count is determined according to the following formula:

for k from 1 to r

bin_{add} = INT((n*k) / (r + 1))

next k

wherein bin_{add} is the sequential bin number to add one to the count of data elements to include therein, n is the total number of bins in the data distribution model, r is the remainder from dividing the number of data elements in the data distribution by the number of bins, and INT is a function for rounding the result of the bracketed formula to produce an integer result.

7. The system of claim 6, wherein the computer program is further for performing the step of counting by counting, for each bin, each data element satisfying the following formula:

start point < element value <= end point

wherein the bin start point is the start point of the respective bin, element value is the value of each data element in each bin, and end point is the end point of the respective bin.

8. The system of claim 7, comprising a storage medium for storing each data distribution model by storing, for each bin, the start point, the end point, the number of data elements, and the parameters of the polynomial formula.

9. The system of claim 1, wherein the computer program is further for performing one or more statistical analysis using the aggregated data model.

10. The system of claim 9, wherein the statistical analysis performed comprises determining the range of the points of the aggregated data model analyzed by subtracting end point of

the last bin in the aggregated data model from the start point of the first bin in the aggregated data model.

11. The system of claim 9, wherein the statistical analysis performed comprises determining the inter quantile range of the points of the aggregated data model.

12. The system of claim 9, wherein the statistical analysis performed comprises determining the median value of the aggregated data model by determining a number j computed by dividing the number of bins by 2, and then reading the value of the end point of the jth bin as the median value if the number of bins in the aggregated data model is equally divisible by 2 or by reading the interpolated value using the polynomial function of the mid point of the jth bin if the number of bins in the aggregated data model is not equally divisible by 2.

13. A method for creating an aggregated data model from a plurality data distribution models, each data distribution model describing a data distribution having one or more data elements, each data element having a value, each data distribution model having one or more bins, each bin comprising a start point having a value, an end point having a value, a value indicating the number of data elements for each bin, and a polynomial formula associated with each bin, the polynomial formula approximating the data elements for the respective bin, the method comprising:

determining which start point has the minimum value and which end point has the maximum value of all of the bins of all of the data distribution models;

setting a start point of a first bin of the aggregated data model to said start point determined to have the minimum value;

setting an end point of a last bin of the aggregated data model to said end point determined to have the maximum value;

determining a total number of a plurality of points for the aggregated data model by adding the values indicating the number of data elements from all bins from all data distribution models;

approximating the data elements in the data distribution described by each data distribution model using the start point, polynomial formula, and number of data elements for each bin in each respective data distribution model, each approximated data element comprising one of said points in the aggregated data model;

sorting the points from minimum to maximum;

distributing the points into one or more bins in the aggregated data model such that a substantially equal number of points are in each bin of the aggregated data model; and

determining a polynomial formula with the sorted data elements for each bin of the aggregated data model.

14. The method of claim 13, comprising determining the end point for each bin in the aggregated data model.

15. The method of claim 13, wherein the step of distributing the points into the one or more bins of the aggregated data model is performed according to the following formula::

(e) if the number of points in the aggregated data model is equally divisible into the number of bins, the end point of the first bin is equal to the value of the ith point in the aggregated data model, wherein i is the number of points in each bin determined by dividing the points equally into the number of bins, wherein the value of the end point of each bin is equal ith point after the last point in the proceeding bin, wherein the start point of each bin is equal to the point after the last point of the previous bin, else

(f) if the number of data elements in the points is not equally divisible by the number of bins, then the number of points in each bin is determined by dividing the number of points by the number of bins, and then adding one to the count of the points in each of a number of bins equal to the remainder after dividing, wherein the bins that have one added to the count is determined according to the following formula:

for k from 1 to r

$$bin_{add} = INT(n*k) / (r + 1))$$

next k

wherein bin_{add} is the sequential bin number to add one to the count of points to include therein, n is the total number of bins in the aggregated

data model, and r is the remainder from dividing the number of points in the data distribution by the number of bins, and INT is a function for rounding the result of the bracketed formula to produce an integer result.

16. The method of claim 13, wherein the steps of approximating the data elements for each bin, determining the end point for each bin, and determining the polynomial formula for each bin are performed separately for each bin of the aggregated data model.

17. The method of claim 13, comprising creating each data distribution model using the following steps:

(f) sorting the data elements in from minimum to maximum for each data distribution;

(g) computing the number of data elements in each data distribution;

(h) determining the value of the start point and the value of the end point of each bin by dividing the data elements into a plurality of substantially equal sized bins for each data distribution;

(i) counting the number of data elements in each bin for each data distribution; and

(j) computing each distribution model for each data distribution, each distribution model comprising, for each bin, the start point of the bin, the end point of the bin, and the number of data elements in the bin.

18. The method of claim 17, comprising determining the start points and end points of the bins for each data distribution model using the following steps:

selecting as the start point of the first bin the value of the data element having the minimum value in the sorted data distribution;

determining the start point and end point of each bin according to the following criteria:

(g) if the number of data elements in the data distribution is equally divisible into the number of bins, the end point of the first bin is equal to the value of the ith data element in the data distribution, wherein i is the number of data elements in each bin determined by dividing the data elements equally into the number of bins, wherein the value of the end point of each bin is equal ith data element after the last data element in the proceeding bin, wherein the start point of each bin is equal to the data element after the last data element of the previous bin, else

(h) if the number of data elements in the data distribution is not equally divisible by the number of bins, then the number of data elements in each bin is determined by dividing the number of data elements by the number of bins, and then adding one to the count of the data elements in each of a number of bins equal to the remainder after dividing, wherein the bins that have one added to the count is determined according to the following formula:

for k from 1 to r

$$bin_{add} = INT(n*k) / (r + 1))$$

next k

wherein bin_{add} is the sequential bin number to add one to the count of data elements to include therein, n is the total number of bins in the data distribution model, r is the remainder from dividing the number of data elements in the data distribution by the number of bins, and INT is a

function for rounding the result of the bracketed formula to produce an integer result.

19. The method of claim 18, wherein the step of counting is performed by counting, for each bin, each data element satisfying the following formula:

start point < element value <= end point

wherein the bin start point is the start point of the respective bin, element value is the value of each data element in each bin, and end point is the end point of the respective bin.

20. The method of claim 19, comprising storing each data distribution model by storing, for each bin, the start point, the end point, the number of data elements, and the parameters of the polynomial formula.

21. The method of claim 13, comprising performing one or more statistical analysis using the aggregated data model.

22. The method of claim 21, wherein the statistical analysis performed comprises determining the range of the points of the aggregated data model analyzed by subtracting end point of the last bin in the aggregated data model from the start point of the first bin in the aggregated data model.

23. The method of claim 21, wherein the statistical analysis performed comprises determining the inter quantile range of the points of the aggregated data model.

24. The method of claim 21, wherein the statistical analysis performed comprises determining the median value of the aggregated data model by determining a number j computed by dividing the number of bins by 2, and then reading the value of the end

point of the jth bin as the median value if the number of bins in the aggregated data model is equally divisible by 2 or by reading the interpolated value using the polynomial function of the mid point of the jth bin if the number of bins in the aggregated data model is not equally divisible by 2.